PATENT

UNITED STATES PATENT APPLICATION

For

A METHOD AND SYSTEM FOR IMPROVING ROBUSTNESS IN A SPEECH SYSTEM

INVENTOR:

LARRY PAUL HECK

Prepared By:
BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP
12400 WILSHIRE BOULEVARD
SEVENTH FLOOR
LOS ANGELES, CA 90025-1026

(408) 720-8300

Attorney Docket No.: 003932.P018

A METHOD AND SYSTEM FOR IMPROVING ROBUSTNESS IN A SPEECH SYSTEM

FIELD OF THE INVENTION

[0001] The present invention relates to the field of speech systems. In particular, the present invention relates to a system and method for improving robustness in a speech system.

BACKGROUND OF THE INVENTION

[0002] Speech and speaker recognition systems are currently in use for responding to various forms of commerce via a voice network. One example of such a system is utilized in conjunction with a stock brokerage. According to this system, a caller can provide his account number, obtain a quotation for the price of a particular stock issue, purchase or sell a particular number of shares at market price or a predetermined target price among other types of transactions.

[0003] The overall performance of such systems are impacted greatly by the environment in which a speaker speaks to the system. For example, speech recognition may be adversely affected in situations when the speaker is speaking in a noisy environment. The noise may include the speech of other people, music, road noise, or any other type of stationary or non-stationary noise. The results of noise may cause the system to misinterpret the commands given by the speaker. In financial industries, the impact of misinterpretations by prior art systems may be disastrous.

SUMMARY OF THE INVENTION

[0004] The present invention introduces a system and method for robust speech recognition by employing a user specific response mechanism in a speech network. In one embodiment, the method comprises receiving an utterance from an intended talker at a speech recognition system. A speaker verification score is computed with a voice characteristic model associated and with the utterance and a speech recognition score associated with the utterance is computed. A best hypothesis associated with the utterance and based on both the speaker verification score and the speech recognition score is selected.

[0005] Other features and advantages of the present invention will be apparent from the accompanying drawings, and from the detailed description, which follows below.

BRIEF DESCRIPTION OF THE DRAWINGS

[0006] The present invention is illustrated by way of example and not intended to be limited by the figures of the accompanying drawings in which like references indicate similar elements and in which:

[0007] **Figure 1** illustrates a speech network according to one embodiment of the present invention;

[0008] **Figure 2** illustrates a computer system representing an integrated multi-processor, in which elements of the present invention may be implemented; and

[0009] **Figure 3** illustrates an exemplary flow diagram of the process performed by an integrated speech and speaker recognizer to provide improved robustness.

DETAILED DESCRIPTION

[0010] A system and method is disclosed for improved robustness in a speech system. In one embodiment, the method comprises receiving an utterance from an intended talker at a speech recognition system. A speaker verification score is computed with a voice characteristic model associated and with the utterance and a speech recognition score associated with the utterance is computed. A best hypothesis associated with the utterance and based on both the speaker verification score and the speech recognition score is selected.

[0011] The present invention also relates to systems for performing the operations, herein. The techniques described herein may be implemented using a general-purpose computer selectively activated or configured by a computer program stored in the computer or elsewhere. Such a computer program may be stored in a computer readable storage medium, such as, any type of disk including floppy disks, optical disks, CD-ROMs, and magnetic-optical disks, read-only memories (ROMs), random access memories (RAMs), EPROMs, EEPROMs, magnetic or optical cards, or any type of media suitable for storing electronic instructions, and each coupled to a computer system bus.

[0012] The algorithms and displays presented herein are not inherently constrained to any particular type of computer or other system. Various general-purpose systems may be used with programs in accordance with the teachings herein, or it may prove convenient to construct more specialized system to perform the required method steps. The required structure for a variety of these systems will be apparent from the description below. In addition, any of a variety

of programming languages, including C++ and Java, may be used to implement the teachings of the techniques described herein.

[0013] Note that in this description, references to "one embodiment" or "an embodiment" mean that the feature being referred to is included in at least one embodiment of the present invention. Further, separate references to "one embodiment" in this description do not necessarily refer to the same embodiment, however, neither are such embodiments mutually exclusive except where so stated or as readily apparent to those skilled in the art.

[0014] **Figure 1** illustrates a speech network according to one embodiment of the present invention. Network 100 may support numerous voice related applications. For example, network 100 may be used to automatically provide a user news, weather, movie times stock quotes, and similar non-secure information. However, network 100 may also be used to automatically provide a user secure information, such as banking information, stock trading information, and similar information that is user specific. Secure information requests are user – specific, in that network 100 should only provide the information if a specified user makes the request. Both secure and non-secure information requests may be generated in network 100. Speech network 100 supports numerous devices for capturing and delivering voice. A user may call into network 100, via a digital telephone 131, used mainly in offices through a call center 130. Traditional, plain old telephone system (POTS) or analog telephones 140 may be used. Cellular telephones 151, via a cellular tower 150 may also be used. Similarly, requests to network 100 may be made through a voice over Internet Protocol (VoIP) microphone 161 via a workstation 160.

[0015] These speech-capturing devices 131-161 may be inter-connected through a wide area network 199, such as the Internet. Voice server 110 receives the information requests and returns a response to the requestor via WAN 199. Voice server 110 provides improved robustness and includes an integrated speaker recognition system and speech recognition system 111 (hereafter "Device 111"). Device 111 includes an automatic speaker recognizer 113 and an automatic speech recognizer 112. Speaker recognizer 113 may also have an integrated speaker detector. The present method and system utilizes the speaker recognizer 113 to improve speech recognizer 112's performance.

[0016] Device 111 is a speaker recognition system that may be a speaker dependent speech recognizer, voice authentication device, and/or speaker identifier. Database 120 is accessible via WAN 199 and stores personal profiles and configuration information, for example, bank account name, address and balance information.

[0017] Situations arise where network 100 must support user specific response and recognition in which improved robustness is beneficial. For example, a user may be driving in automobile 152, and he attempts to place an order to sell 100 shares of company ABC stock. However, just as the user is about to utter, "Sell 100 ABC," his son from the back seat screams, "Sell 100 XYZ." The present method and system for improved robustness in speech network 100 may continue to track the intended speaker and recognize that it was not the desired speaker who spoke, and reject the son's utterance, instead of placing a false order.

[0018] Another voice application where the network 100 supports user specific response and recognition occurs when two users are speaking at the same time. Known prior art systems perform recognition only on the first user, however, if the first user's speech pauses, the prior art system catches on to the second user and recognizes the speech of the second user, believing the second user is still the first user.

[0019] Thus, the present method and system may be used to improve speech recognition performance, or specifically robustness in cases of interfering speakers by using information about the user's voice. Furthermore, the present method and system may be used to improve speech recognition in noisy environments that do not only include interfering talkers. For example, elements of network 100 could be used in conjunction with a hands-free microphone kit for a cellular telephone, inside an automobile. In addition to multiple speakers, there may be noise from the highway and engine. The present method and system may also be used in noisy cafes, subways, trains, and crowds where improved robustness is desired.

[0020] In one embodiment, a voice application arises where a single unidirectional microphone is placed near the desired speaker, although systems having multiple or omni-directional microphones may also be used to add the element of space to the system. The single microphone embodiment adds additional challenges due to the fact that all noises, including the desired speaker are mixed at the microphone. Thus, the separation techniques for isolating the desired speech from the other noise, must be done electronically using energy and timing information instead of space or range information.

[0021] **Figure 2** illustrates a computer system representing an integrated multi-processor, in which elements of the present invention may be implemented. Computer system 200 may represent the hardware architecture of voice server 110 that includes device 111.

[0022] Computer system 200 may also represent workstations 160, or call center 130. In one embodiment of the present invention, computer system 200 may be an IBM® compatible personal computer (PC), an Apple Macintosh® personal computer, or a SUN® SPARC workstation. The operating system run on computer system 200 may be Windows®, LINUX, SUN® Solaris® or other similar operating system.

[0023] One embodiment of computer system 200 comprises a system bus 220 for communicating information, and a processor 210 coupled to bus 220 for processing information. Computer system 200 further comprises a random access memory (RAM) or other dynamic storage device 225 (referred to herein as main memory), coupled to bus 220 for storing information and instructions to be executed by processor 210. Main memory 225 may be used for storing temporary variables or other intermediate information during execution of instructions by processor 210. Computer system 200 also may include a read only memory (ROM) and/or other static storage device 226 coupled to bus 220 for storing static information and instructions used by processor 210.

[0024] A data storage device 227 such as a magnetic disk or optical disc and its corresponding drive may also be coupled to computer system 200 for storing information and instructions. Computer system 200 may also be coupled to a second I/O bus 250 via an I/O interface 230. A plurality of I/O devices may

be coupled to I/O bus 250, including a display device 243, an input device (e.g., an alphanumeric input device 242 and/or a cursor control device 241). For example, video news clips and related information may be presented to the user on the display device 243.

[0025] The communication device 240 is for accessing other computers (servers or clients) via a network. The communication device 240 may comprise a modem, a network interface card, or other well-known interface device, such as those used for coupling to Ethernet, token ring, or other types of networks.

[0026] **Figure 3** illustrates an exemplary flow diagram of the process 300 performed by device 111 to provide improved robustness. The process begins at block 301. At processing block 305, device 111 detects the desired or intended talker using speaker recognizer 113. The environment surrounding the intended talker may be noisy, or silent. The intended talker may be registered with network 100 - that is a voice characteristic model may already be stored in database 120 for the intended talker. In another embodiment, the intended talker has no pre-existing voice model stored on network 100. The detection may occur by having the intended talker provide device 111 an identity claim to authenticate and verify. In another embodiment, device 111 may select a dominant voice as the intended talker in a noisy environment by measuring the energy levels associated with the speech received by device 111. In yet another embodiment, the detection may occur by using hotword speech recognition. For example, device 111 may detect the intended talker, when the intended talker says the hotword "computer."

[0027] Flow continues to processing block 310, where an initial model of the intended talker's voice characteristics is created. The model may be a voice print, or any data that provides a personal profile such as acoustic and/or linguistic characteristics unique to the intended talker.

[0028] At processing block 315, device 111 receives a second additional utterance or utterances from the intended talker who is in the noisy environment. The noisy environment includes other speakers talking in the background, music, highway road sounds and any other stationary or non-stationary noise. The first and second utterances may be two portions of the same sentence or of the same word. Flow continues to processing block 317 where a portion of the second utterance is processed by device 111. This portion of the second utterance may be the entire utterance received in block 315, or it may be a word spoken in the utterance, a phonetic syllable ("phoneme") of the word, or a single frame. In the case of processing an individual frame, the frame may be a few milliseconds, for example in one embodiment of the invention, the frame may be 30 milliseconds of speech.

[0029] Processing flow continues to processing blocks 320 and 325. At processing block 320, the speaker recognizer 113 generates a speaker verification score for the intended talker. While the speaker recognizer 113 generates the speaker verification score, speech recognizer 112 interprets the portion of the second utterance, and generates a speech recognition score. The speech recognition score is indicative of speech recognizer 112's confidence in determining that the portion of the second utterance was correctly processed.

[0030] Flow continues to processing block 330, where the speaker verification and speech recognition scores are combined. The scores may be combined in numerous ways, with different weightings. In one embodiment, the phoneme- and state-dependent speech recognition scores for a specific frame of data are scaled by a weight between 0 and 1, depending on the verification score for that frame. Specifically, if the verification score is very high (i.e., there is a very good match between the data and the voiceprint), then the recognition score for the particular phoneme/state is weighted by a scalar close to 1 (unaltered), whereas if the verification score is very low (i.e., the data does not match the voiceprint), then the recognition score is scaled by a very small weight close to zero. In this way, if the processed data originated from a sound source other than the intended talker, then the verification score will be low and the speech recognition system's highest-scoring string of phonemes/words (search path) up to that point in the search will not be substantially altered by the data.

[0031] At processing block 340 speech recognition search paths are updated in speech recognizer 112 based on the combined score to generate a list of best hypothesis of what the intended talker is saying. The score may be altered based on a forward Viterbi search (left-to-right, i.e. time=0 until the time when the utterance is complete), such that alterations are made at the sentence, word, phonetic or frame level. In alternate embodiments, the search may be forward, backward or multipass.

[0032] Flow continues to decision block 345 where device 111 determines if the last portion of speech in the utterance has been processed. If the last portion of speech has not been processed, then flow continues to processing

block 346 where the next portion of speech is called and flow is passed back to processing block 317. If the last portion of speech has been processed, then flow continues to block 350. At processing block 350, the best hypothesis is selected and used by the network 100 to generate a response. The process ends at block 399.

[0033] On a frame level, speech recognizer 112 (in block 325) performs a computation that determines the likelihood that the frame of speech comes from a particular phoneme. For example, recognizer 112 may compute the likelihood that a 30 millisecond frame of speech came from the phoneme "ah." In addition to performing the computation with the data of the given frame, the likelihood computation also takes into account previous likelihood computations and information about the next frame that will be processed.

[0034] In the foregoing specification, the invention has been described with reference to specific exemplary embodiments thereof. It will, however, be evident that various modifications and changes may be made thereto without departing from the broader spirit and scope of the invention as set forth in the appended claims. The specification and drawings are, accordingly, to be regarded in an illustrative sense rather than a restrictive sense.